

Enhancing In-the-Wild Speech Emotion Conversion with Resynthesis-based Duration Modeling

Navin Raj Prabhu, Danilo de Oliveira
Signal Processing
University of Hamburg
 Hamburg, Germany
 {firstname.lastname}@uni-hamburg.de

Nale Lehmann-Willenbrock
Industrial and Organizational Psychology
University of Hamburg
 Hamburg, Germany
 nale.lehmann-willenbrock@uni-hamburg.de

Timo Gerkmann
Signal Processing
University of Hamburg
 Hamburg, Germany
 timo.gerkmann@uni-hamburg.de

Abstract—Speech Emotion Conversion aims to modify the emotion expressed in input speech while preserving lexical content and speaker identity. Recently, generative modeling approaches have shown promising results in changing local acoustic properties such as fundamental frequency, spectral envelope and energy, but often lack the ability to control the duration of sounds. To address this, we propose a duration modeling framework using resynthesis-based discrete content representations, enabling modification of speech duration to reflect target emotions and achieve controllable speech rates without using parallel data. Experimental results reveal that the inclusion of the proposed duration modeling framework significantly enhances emotional expressiveness, in the in-the-wild MSP-Podcast dataset. Analyses show that low-arousal emotions correlate with longer durations and slower speech rates, while high-arousal emotions produce shorter, faster speech.

Index Terms—Speech emotion conversion, duration modeling, non-parallel samples, arousal, in-the-wild

I. INTRODUCTION

Speech is a fundamental social signal that plays a key role in enabling interactions, whether between humans or between humans and machines. It conveys essential information for the interaction, including lexical content, speaker identity, and expressed emotions [1]. The task of speech generation and synthesis thereby is a crucial research topic in the fields of signal processing and human-computer interaction. With the advent of generative deep neural networks, substantial improvements have been made in speech generation and synthesis [2]–[5]. However, emotion-conditioned speech synthesis remains a significant challenge [6]–[9]. In the context of human-computer interaction, the need for emotion-conditioned speech synthesis is crucial, not only to improve the naturalness and expressiveness of machine communication but also to enhance user engagement, foster empathy, and enable more effective and context-aware responses [9], [10].

Speech Emotion Conversion (SEC) is a sub-field of emotion-conditioned speech synthesis that aims to modify the emotion expressed in input speech while preserving lexical content and speaker identity [6], [8], [11]. This requires precise control over prosodic attributes that convey emotional content,

such as intonation, stress, rhythm, and loudness, which are controlled by the acoustic features of speech sounds, such as fundamental frequency, duration, energy, and spectral envelope. While it is appealing to control these attributes based on a target emotion, changing the corresponding acoustic feature for each prosodic component presents its own unique set of challenges [9].

Generative deep neural networks, such as variational autoencoders (VAEs) [12], generative adversarial networks (GANs) [13], and diffusion models [14], have been employed to the task of SEC with success in emotion conversion capabilities and improved naturalness in generated speech [10], [15]. However, these methods often overlook duration modeling in emotion conversion, resulting in inadequate control over crucial prosodic features such as rhythm and stress. Instead, they typically enforce fixed durations, where the emotion-converted output speech sounds have exactly the same duration as in the input, regardless of the intended emotion change. Interestingly, this is in contrast to the task of text-to-speech (TTS) synthesis, where duration modeling with duration-flexible speech generation is a common module, with proven improvements in the naturalness of synthesised speech [4].

Durflex-EVC [9] introduced duration modeling in SEC with parallel data, where for each source utterance with a corresponding source emotion also a corresponding target utterance with a target emotion is available. Durflex-EVC learns discrete speech units from parallel target emotion speech and their repetitions. However, a particular challenge in duration modeling for emotion conversion arises when working with in-the-wild emotion datasets, as these lack parallel samples. As a result, there is no ground-truth duration reference for the target emotion, making accurate duration control more challenging. While in-the-wild datasets offer a richer and more naturalistic collection of emotional speech, along with greater speaker diversity and varied acoustic conditions [16]–[18], their non-parallel nature limits their applicability for supervised duration modeling in emotion conversion [10], [15]. In this work, we aim to achieve duration modeling in SEC, focusing specifically on in-the-wild datasets without relying on parallel data.

In this work, we propose a resynthesis-based duration modeling approach to enhance SEC performance, which operates on discrete speech units and does not require parallel data.

This work was funded under the Excellence Strategy of the Federal Government and the Länder, and the project “Mechanisms of Change in Dynamic Social Interaction” (LFF-FV79, Landesforschungsförderung Hamburg).

To enable duration modeling in a non-parallel setting, the proposed method is trained using a resynthesis setup, inspired by [10] and [15]. In this setup, during training, the model simultaneously reconstructs the original input speech while the duration model learns to predict the repetition counts of discrete speech units. This prediction is based solely on the input speech, without any reliance on target speech. During inference, the trained duration model can predict the appropriate unit repetitions based on a target emotion embedding, enabling emotion-aware duration control. Experiments in the in-the-wild MSP-Podcast dataset show that the inclusion of the proposed duration modeling framework is beneficial for emotional expressiveness.

II. RELATED LITERATURE

A. Speech Emotion Conversion techniques

SEC techniques can be broadly categorized into *sequential* speech generation and *parallel* speech generation models. Sequential generation models (e.g., [19]–[21]) perform emotion-conditioned speech synthesis by sequentially generating speech units or frames, thereby achieving implicit duration modeling. However, they often face challenges such as difficulty in capturing long-term dependencies and high time complexity [9]. This has motivated the development of parallel generation models (e.g., [7], [9], [22]), which address these limitations by enabling parallel generation of speech frames. However, a key requirement of these models is the explicit modeling of the intended duration [9].

Recently, there has been a shift in voice and emotion conversion research away from traditional scripted or acted-out speech, which often lacks the natural spontaneity of real-life conversations, towards the use of in-the-wild recorded speech [10], [15], [16]. Unlike acted-out data, which is essentially read-out speech, in-the-wild recordings are more spontaneous and capture diverse speaking styles, emotional expressions, nonverbal cues like laughter and lip smacks, and disfluencies such as repetitions, hesitations, and interruptions [16], [23]. Empirical analyses using the NaturalVoices dataset [16] show that models trained on in-the-wild samples generate more natural and intelligible speech. However, such training requires methods that do not depend on parallel speech samples.

Raj Prabhu et al. [10] proposed a SEC framework using *resynthesis* to eliminate the need for parallel data. A HiFiGAN-based vocoder reconstructs input speech from disentangled self-supervised learning (SSL)-based representations: discrete HuBERT embeddings for lexical content, speech emotion recognition (SER)-derived emotion embeddings, and speaker verification-based speaker embeddings. At inference, modifying the emotion embedding enables synthesis with the target emotion. Building on this, EmoConv-Diff [15] uses a diffusion decoder conditioned on “average-phoneme” mel features. While effective for in-the-wild SEC, these approaches lack duration modeling and cannot control speech rate based on the target emotion.

In this work, we use the resynthesis technique to achieve duration modeling under in-the-wild conditions without relying

on any information from target emotion speech, neither target emotion durations nor speech embeddings. To the best of our knowledge, this is the first study to propose duration-flexible SEC that does not rely on parallel emotion speech samples.

B. Duration Modeling techniques

Duration Modeling in speech synthesis has been approached as a task of predicting the temporal alignment between lexical tokens (e.g., phonemes) and their respective acoustic features, essentially determining how long each unit should be held in the synthesised speech [4], [24]. Modern neural TTS systems incorporate duration modeling either implicitly, using the attention mechanism [25], [26], or explicitly, using a duration predictor that predicts phoneme repetitions [4], [24], [27]. The explicit modeling approach has been preferred in non-autoregressive models like Grad-TTS [4] and FastSpeech [24], allowing for greater flexibility in modifying speaking style, emphasis, or speech rate.

Despite its demonstrated effectiveness in TTS, duration modeling has received limited attention in tasks like emotion and voice conversion. A likely reason for this omission is the difficulty of jointly training a duration model and learning to modify the prosodic features of the input speech during conversion. As a result, many emotion conversion models adopt a fixed-duration strategy, where the converted speech maintains the same duration as the input, regardless of the target emotion [10], [15]. This constraint limits the expressiveness of SEC systems by restricting their ability to adjust the timing of lexical units, and consequently, the rhythm and speech rate aligned with the intended emotional state.

DurFlex-EVC [9] addresses the gap of incorporating duration modeling in SEC by using a so-called *Unit Aligner* module to extract discrete content tokens and a *Duration Predictor* to estimate their repetitions. However, this approach is not directly applicable to in-the-wild datasets, as it relies on speech units extracted from parallel target speech, which are unavailable in non-parallel settings. Additionally, the use of look-up table-based speaker and emotion embeddings further limits its adaptability to in-the-wild scenarios, where speaker and emotion conditions are more variable and less structured. Similarly, [28] and [29] also address duration modeling. While [28] target speaker conversion and [29] focus on emotion conversion with acted data and categorical emotion labels.

In this work, inspired by [10] and [15], we propose a resynthesis-based duration modeling approach that is better suited for in-the-wild datasets. Our method relies solely on the input speech during training and does not require any information from the target emotion or target speech, making it fully compatible with non-parallel SEC tasks.

III. METHODOLOGY

The overall task of speech emotion conversion can be formulated as follows: given a single-channel audio input $\mathbf{x}_{l,s,e} \in \mathbb{R}^{1 \times T}$ representing a spoken utterance with lexical content l , speaker identity s , and annotated emotion level e , where the raw waveform is denoted as a sequence of samples

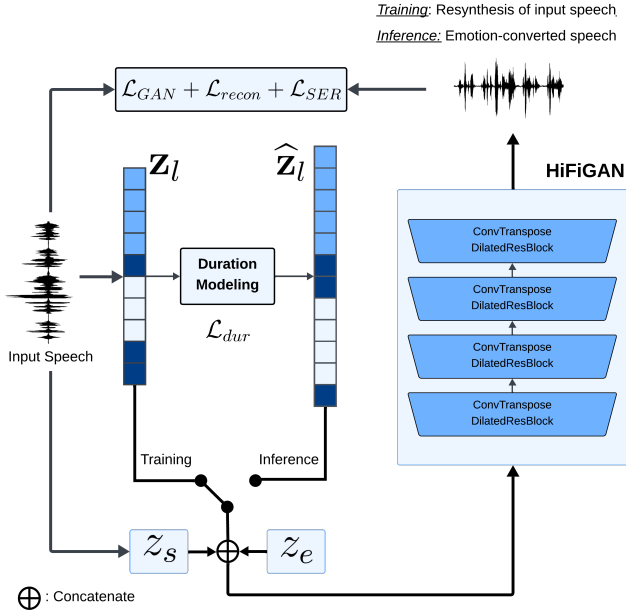


Fig. 1: Overview of the speech emotion conversion framework.

$\mathbf{x} = [x_1, \dots, x_T]$, the goal is to generate $\hat{\mathbf{y}}_{l,s,\bar{e}} \in \mathbb{R}^{1 \times T'}$, with T' potentially different from T . This output should preserve the original lexical content l and speaker identity s from $\mathbf{x}_{l,s,e}$, while converting the expressed emotion to a desired target level \bar{e} . With that intent, the length T' is jointly modeled and the generated output is expected to be duration-flexible with respect to the desired target emotion \bar{e} . We adopt the SSL-based HiFiGAN model from [10] as the SEC backbone for integrating our resynthesis-based duration modeling approach. Its original design, which is also trained using a resynthesis paradigm, makes it particularly well-suited for this purpose. The overall SEC methodology is depicted in Figure 1.

A. Disentangled Representations

For the disentangled SSL-based representations input to the HiFiGAN decoder, we use the following encoded features:

- (i) *Lexical representation* ($\mathbf{z}_l \in \mathbb{N}^{1 \times N}$): Following [9]–[11], we use discrete HuBERT units obtained via k -means clustering on continuous HuBERT features. Formally, $\mathbf{z}_l = [z_1, \dots, z_N]$, where each z_i is a positive integer and N is the length of the input discrete unit sequence, corresponding to the number of frames in HuBERT’s representations. Prior studies [30]–[32] have shown that these units strongly correlate with the phonemic content of the utterance. The feature rate of these speech units is 49Hz.
- (ii) *Speaker representation* ($\mathbf{z}_s \in \mathbb{R}^{512}$): Adopted from [3], we use a d -vector extracted from a pretrained WavLM-based speaker verification model [33].
- (iii) *Emotion representation* ($\mathbf{z}_e \in \mathbb{R}^{128}$): A continuous embedding obtained by applying a trainable linear transformation to the emotion label e during training, and to the target emotion label \bar{e} during inference.

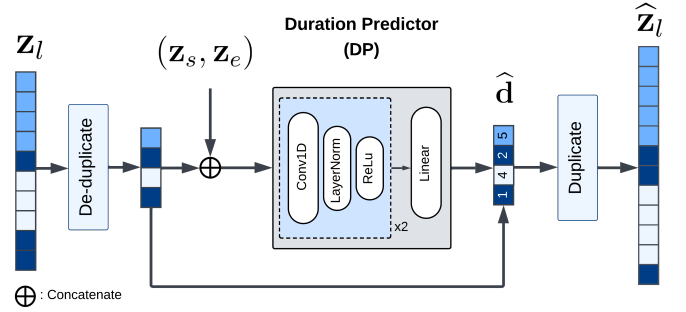


Fig. 2: Overview of the duration modeling technique.

Unlike \mathbf{z}_l , both \mathbf{z}_s and \mathbf{z}_e are global utterance-level representations. To align them with the frame-level \mathbf{z}_l , we broadcast \mathbf{z}_s and \mathbf{z}_e across frames/discrete units, resulting in \mathbf{z}_s and \mathbf{z}_e .

B. Duration Modeling

An overview of the duration modeling technique can be seen Figure 2. Based on \mathbf{z}_s and \mathbf{z}_e , we perform duration modeling on the discrete HuBERT speech units \mathbf{z}_l , which represent the lexical content of input speech. Formally, we formulate the resynthesis-based duration modeling as follows: for \mathbf{z}_l of input speech, we train a *Duration Predictor* (DP) to predict the consecutive repetition of discrete speech units \mathbf{d} , conditioned on emotion and speaker representations. These repetitions represent the durations of each lexical unit.

Firstly, the frame-level \mathbf{z}_l is de-duplicated to extract unit-level speech units, where the repetitions are ignored to obtain consecutive unique speech units. Secondly, this unit-level representation is fed as input to the predictor DP. To further achieve speaker and emotion conditioned duration modeling, we concatenate the speaker and emotion embeddings ($\mathbf{z}_s, \mathbf{z}_e$) and pass them as an additional input to the predictor.

The predictor is a simple deterministic neural network comprising two convolution layers and a linear layer to predict \mathbf{d}_i for respective unit-level speech units, where i is the index in the de-duplicated sequence of discrete speech units. As an example, if the speech units in \mathbf{z}_l are $[1, 1, 2, 2, 2, 1, 3, 3, 3, 3]$, the de-duplicated sequence would be $[1, 2, 1, 3]$, and the target \mathbf{d} would be $[2, 3, 1, 4]$. For stable training and to better account for outliers in durations, we predict durations in the log-scale: $\log(\mathbf{d})$, as suggested in [4]. During training, the predicted log-scale durations/repetitions are directly used in the loss function and the true frame-level \mathbf{z}_l is used as the input to the HiFiGAN decoder. However, during inference, the predicted log durations $\widehat{\log(\mathbf{d})}$ are reversed back into duration units as follows:

$$\hat{\mathbf{d}} = \min \left(1, e^{\widehat{\log(\mathbf{d})} + 1} \right). \quad (1)$$

Finally, the reversed durations $\hat{\mathbf{d}}$ are used to duplicate the unit-level speech units to obtain the duration modeled discrete lexical units $\hat{\mathbf{z}}_l$. Note that, as per the resynthesis paradigm, we use the estimated $\hat{\mathbf{z}}_l$ only during inference, and during training the true \mathbf{z}_l is used. The final input to the HiFiGAN decoder is

the combined concatenated representation: $(\mathbf{z}_l, \mathbf{z}_s, \mathbf{z}_e)$ during training and $(\hat{\mathbf{z}}_l, \mathbf{z}_s, \mathbf{z}_e)$ at inference time.

C. Loss Functions

The overall training of the SEC architecture involves four different loss terms: (i) the adversarial based HiFiGAN loss \mathcal{L}_{GAN} , which is the same as used in [11] and [10], (ii) a reconstruction loss,

$$\mathcal{L}_{recon}(G) = \sum_{\mathbf{x}} \|\phi(\mathbf{x}) - \phi(\hat{\mathbf{y}})\|_1, \quad (2)$$

where ϕ is a function computing Mel-spectrogram, (iii) a speech emotion recognition loss which is used to better condition the SEC model on the emotion of input speech,

$$\mathcal{L}_{SER} = \sum_{\mathbf{x}} [1 - L_{ccc}(e, E_{SER}(\hat{\mathbf{y}}))], \quad (3)$$

where L_{ccc} is the concordance correlation coefficient (CCC) [34] computed between the ground-truth emotion e of input speech, and the predicted emotion for resynthesised speech $E_{SER}(\hat{\mathbf{y}})$, and finally, (iv) the duration modeling loss \mathcal{L}_{dur} . We use the speech emotion recognition (SER) model introduced in [35] as the emotion predictor $E_{SER}(\cdot)$. The emotion predictor is a wav2vec-based neural network trained on the MSP-Podcast dataset to predict the arousal of input speech.

As the duration modeling loss \mathcal{L}_{dur} , we experiment with three different loss functions, all computed on the logarithm of the ground-truth durations. Let the predicted log-durations be $\hat{\mathbf{d}} = (\hat{d}_1, \hat{d}_2, \dots, \hat{d}_U)$, and the ground-truth durations be $\mathbf{d} = (d_1, d_2, \dots, d_U)$, where $\tilde{d}_u = \log d_u$. Specifically, the four loss functions are: (i) *mean squared error* (\mathcal{L}_{mse}), (ii) *mean absolute error* (\mathcal{L}_{abs}), and (iii) *uncertainty-based negative log-likelihood* (NLL) Loss, assuming a Gaussian distribution over log durations with predicted mean \hat{d}_u and predicted standard deviation σ_u (\mathcal{L}_{NLL}).

Finally, the overall speech emotion conversion loss of the architecture is as follows:

$$\mathcal{L}_{SEC} = \lambda_1 \mathcal{L}_{GAN} + \lambda_2 \mathcal{L}_{recon} + \lambda_3 \mathcal{L}_{SER} + \lambda_4 \mathcal{L}_{dur}, \quad (4)$$

where values of λ_1 , λ_2 and λ_3 are adopted from [10], and λ_4 is set to 2 after a grid-search based tuning.

IV. EXPERIMENTAL SETUP

A. Dataset

The dataset used in this study is the *in-the-wild* MSP-Podcast dataset (v1.10) [23], which contains approximately ≈ 238 hrs of audio sourced from podcasts, with utterance-level emotion annotations provided in terms of arousal, valence, and dominance. The dataset in contrast to predominant SEC datasets (e.g., ESD [8], IEMOCAP [36]) is larger, has utterances of variable duration, has over 1400 speakers, and contains naturalistic emotional expressions. For example, the ESD contains acted-out utterances from only 10 English speakers and only ≈ 29 hours of acted-out utterances. To the best of our knowledge, this is one of the few works to perform SEC on an in-the-wild dataset, along with [10] and [15].

Model	DP	WVMOS \uparrow	SER Error \downarrow	
			L_{mse}	L_{abs}
HiFiGAN [10]	\times	3.26	0.084	24%
EmoConv-Diff [15]	\times	2.56	0.072	21%
MSE	\checkmark	3.42	0.072	21%
L_1	\checkmark	3.36	0.075	22%
+UnitAligner	\checkmark	3.16	0.086	27%
Uncert	\checkmark	3.30	0.069	20%

TABLE I: Overall performance of model versions. DP: Duration Predictor, \checkmark indicates the respective model includes duration modeling, and \times indicates it's absence.

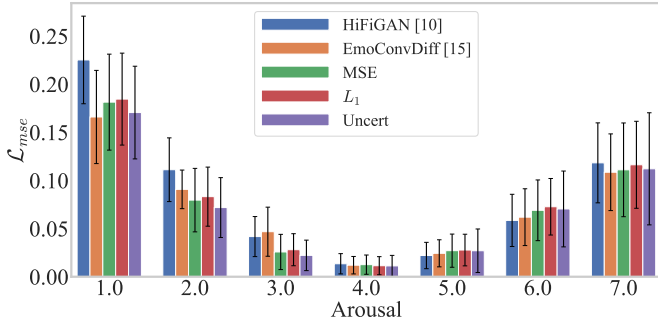
For the purpose of this work, we focus exclusively on arousal annotations for SEC, following prior works on SEC under in-the-wild conditions [3], [15]. Performing SEC on the arousal dimension, instead of categorical representation has two advantages: (i) the circumplex-model based representation better captures the subtle difference between human emotion categories [37], [38], and (ii) achieve implicit intensity control [10], as opposed to an additional effort in the categorical representation case. The arousal annotations are rated on a 1–7 scale and exhibit a distribution with a mean $\mu = 4$ and standard deviation $\sigma = 0.95$. This indicates that samples are more concentrated in the mid-range (scores 3 to 5), with fewer examples at the extremes (scores 1 and 7). This skewed distribution mirrors the nature of emotional expression in real-world, in-the-wild scenarios, such as podcasts, where extreme emotional states are relatively rare.

B. Validation Measures

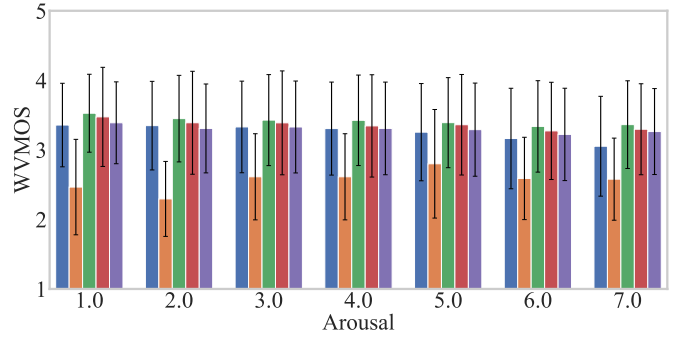
We evaluate the proposed methodology based on two key aspects: its speech emotion conversion (SEC) capabilities and the naturalness of the synthesised speech. To assess SEC performance, we use mean-squared error (\mathcal{L}_{mse}) and mean-absolute error (\mathcal{L}_{abs}), both computed between the target arousal \bar{e} and the SER model's prediction on the resynthesised output, $E_{SER}(\hat{\mathbf{y}})$. For measuring the naturalness of the synthesised speech $\hat{\mathbf{y}}$, we employ the wav2vec mean-opinion score (WVMOS) [39], an objective speech quality metric derived from wav2vec2.0 [40]. WVMOS is fine-tuned on mean-opinion scores (ranging from 1 to 5) collected through listening tests from the 2018 Voice Conversion Challenge [41], which focused specifically on naturalness. This makes WVMOS a suitable non-intrusive metric for evaluating the naturalness of $\hat{\mathbf{y}}$. It's important to note that, since we do not use parallel data, we rely solely on non-intrusive evaluation metrics that do not require access to the ground-truth audio $\mathbf{y}_{l,s,\bar{e}}$ for emotion conversion.

C. Model Versions

As baselines for performance comparison, we use the HiFiGAN [10] and EmoConv-Diff [15] architectures introduced earlier. Both are designed to handle in-the-wild data using the resynthesis paradigm that does not rely on parallel samples, similar to our proposed approach. Notably, neither model includes duration modeling, making them appropriate baselines



(a) SEC performance ($\mathcal{L}_{mse} \downarrow$).



(b) WVMOS \uparrow performance.

Fig. 3: Speech emotion conversion performance and naturalness of generated speech across arousal levels.

for evaluating its impact. In fact, the HiFi-GAN architecture also serves as the backbone for our SEC model, into which we integrate duration modeling, further justifying its role as a baseline. We evaluate four variants of the proposed model, each employing a different approach to duration modeling,

- (i) *MSE*: trains the duration predictor using mean squared error (MSE) loss.
- (ii) *L_1* : replaces MSE with mean absolute error.
- (iii) *+UnitAligner*: integrates the Unit Aligner module from [9], which learns discrete speech units directly from data instead of relying on pretrained HuBERT units. These learned units are then utilized by the duration predictor, improving alignment between units and acoustic frames. With the inclusion of the Unit Aligner, this baseline corresponds to a reimplement of DurFlex [9] in our non-parallel, in-the-wild setting.
- (iv) *Uncert*: introduces an uncertainty-aware duration predictor that estimates both the mean and variance of durations and is trained using Negative Log-Likelihood (NLL) loss for a probabilistic formulation.

V. RESULTS

A. Influence of Duration Modeling

The overall performance of the different versions of the proposed model, as compared to the baselines, is shown in Table I. From the results, we observe the following: Firstly, incorporating duration modeling into the HiFi-GAN baseline leads to both increased naturalness in generated speech and enhanced speech emotion conversion capabilities. The MSE variant of the duration modeling attains a WVMOS of 3.42 and a \mathcal{L}_{abs} of 21%, representing an improvement over the HiFi-GAN baseline, which achieves a WVMOS of 3.26 and a \mathcal{L}_{abs} of 24%. Secondly, it is evident that, except for the +UnitAligner version, all other duration modeling approaches consistently outperform the HiFi-GAN baseline, highlighting the significance of duration modeling for SEC. A probable reason why the UnitAligner does not contribute to improved duration modeling is that it is better suited for training scenarios where parallel data samples are available, as was the case in the work that originally introduced it [9], and it does

not provide additional benefit in a resynthesis-based training paradigm, where direct usage of HuBERT speech units \mathbf{z}_l without alignment is more appropriate. Thirdly, we note that while duration modeling yields a considerable improvement over the HiFi-GAN baseline, the gains over EmoConv-Diff are relatively small. This could potentially be attributed to differences in the decoder itself, as the diffusion-based decoder used by EmoConv-Diff is more complex and has already demonstrated improvements over HiFi-GAN decoders in TTS tasks [4]. Finally, among the duration modeling variants, the *MSE* and *Uncert* approaches emerge as the most effective. The *MSE* variant yields slightly better naturalness, while the *Uncert* variant achieves marginally better SEC performance. Overall, based on the empirical results, we recommend the *Uncert* variant for duration modeling due to its strong SEC performance and improved naturalness over the baseline. The SEC capabilities can be further noted in the audio examples presented online¹.

B. Performances across arousal levels

In Figure 3, the performance results are illustrated according to the target arousal level of the emotion-converted speech, considering both SEC capabilities (Fig. 3a) and the naturalness of the generated speech (Fig. 3b). Regarding SEC performance, the results in Fig. 3a indicate that incorporating duration modeling proves particularly advantageous for generating low-arousal speech, with the duration modeling variants showing a noticeably larger improvement over the baseline at low arousal levels, and only a slight improvement at high arousal. Additionally, we observe that EmoConv-Diff achieves the best SEC performance for the extreme target arousal levels of 1 and 7, with the Uncert variant of duration modeling coming closest in performance.

From Fig. 3b, we observe that duration modeling approaches consistently yield more natural-sounding speech compared to both the HiFi-GAN and EmoConv-Diff baselines. This underscores the importance of explicit duration modeling for enhancing speech naturalness. Although EmoConv-Diff demonstrates competitive SEC performance, it notably lacks

¹<https://sp-uhh.github.io/emoconv-gen>

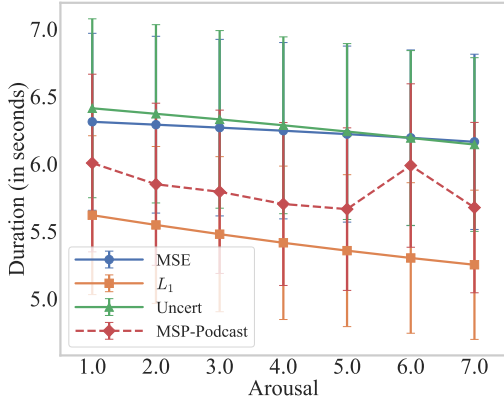


Fig. 4: Mean and standard deviation of durations for emotion-converted speech across target arousal levels. The red dashed line represents the durations from the MSP-Podcast dataset.

in naturalness. This suggests a promising direction for future work: incorporating duration modeling into diffusion-based SEC methods such as EmoConv-Diff.

C. Duration-flexible speech emotion conversion

To evaluate the effectiveness of duration modeling, Figure 4 presents the mean durations (in seconds) and their standard deviations for emotion-converted speech generated by the various model variants. Additionally, we include the oracle mean durations from the MSP-Podcast dataset (denoted by the red dashed line) corresponding to the ground-truth arousal levels. The figure clearly illustrates that all duration modeling variants successfully capture the inverse linear relationship between arousal level and speech duration: models tend to generate longer speech for low arousal levels and shorter speech for high arousal levels. This trend, also evident in the dataset reference line, is well reproduced by the SEC models incorporating duration modeling.

Among the different variants, the L_1 loss shows the most pronounced duration contrast, with the largest difference in mean duration between arousal level 1 and arousal level 7, measured as $\Delta_{1-7} = 0.37$ secs. Both the *MSE* and *Uncert* variants reflect similar patterns, with the *Uncert* variant yielding a slightly higher Δ_{1-7} of 0.21 secs. Overall, the L_1 variant tends to generate shorter duration speech compared to the other models. This behavior may stem from the nature of L_1 loss, being based on absolute error, is less sensitive to outliers (e.g., highly repetitive speech units). Consequently, it may underfit to high-repetition segments, treating them as noise and favoring shorter durations in general.

D. Modification of prosody features

To examine the prosodic modifications achieved by the duration modeling-based SEC architecture, we present Figure 5. It shows the pitch contours of the input speech (represented by black dashed lines), alongside the emotion-converted speech for target arousal level 1 (extreme low arousal, shown in blue) and arousal level 7 (extreme high arousal, shown in red). In

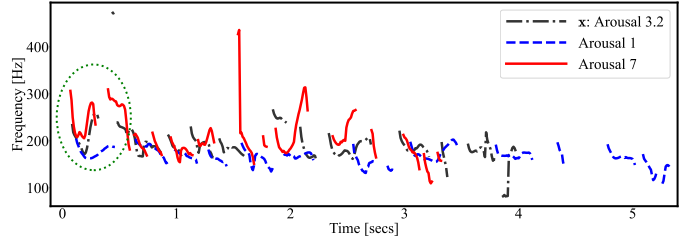


Fig. 5: Pitch contours of the input speech X , along with emotion-converted speech for target arousal levels 1 and 7.

the figure, a green dotted ellipse is used to highlight a region of interest. Due to the strong duration control demonstrated by the L_1 variant, as shown in Sec. V-C, the pitch contour analysis is conducted on speech generated by this variant.

The pitch contours in the figure reveal that the synthesised speech for high arousal ($\bar{e} = 7$) exhibits a higher mean pitch and greater pitch variability compared to both the ground-truth speech ($\bar{e} = 3.20$) and the synthesised speech for low arousal ($\bar{e} = 1$). This observation is consistent with prior studies linking high emotional intensity to increased mean pitch [21], and aligns with baseline research demonstrating effective pitch control. More importantly, we observe the impact of duration modeling, which yields a shorter duration for high arousal speech (≈ 3.5 secs), and a longer duration for low arousal speech (≈ 5.3 secs), compared to the ground-truth input speech of mid-level arousal (≈ 4 secs). Finally, within the highlighted region of interest (indicated by green dotted lines), it is evident that duration modeling enables effective control and modification of speech rate. Specifically, the high arousal speech, while exhibiting a higher mean and variability in pitch, also features a noticeably shorter voiced segment (red contour) than the corresponding voiced segment in the low arousal speech (blue contour).

VI. CONCLUSION

In this work, we proposed a resynthesis-based duration modeling approach for speech emotion conversion that does not require parallel target speech samples—a key challenge due to the unavailability of ground-truth lexical durations during training. To overcome this, we employed a resynthesis training paradigm where the model learns to reconstruct input speech conditioned on lexical, emotion, speaker, and duration information. At inference time, emotion conversion is achieved by modifying the emotion embeddings.

We validated our approach on an in-the-wild dataset, evaluating both emotion conversion accuracy (using a pretrained SER model) and the naturalness of synthesised speech (via WVMOS). Pitch contour analysis confirms that our approach achieves not only pitch modulation but also speech rate control, producing shorter, faster speech for high arousal and longer, slower speech for low arousal. The results demonstrate the effectiveness of duration modeling, with consistent improvements in both SEC performance and naturalness over baseline methods.

REFERENCES

- [1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," vol. 61, no. 5. ACM New York, NY, USA, 2018, pp. 90–99.
- [2] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.
- [3] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, 2020.
- [4] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *Int. Conf. Machine Learning (ICML)*. PMLR, 2021.
- [5] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim, "Diff-tts: A denoising diffusion model for text-to-speech," *arXiv preprint arXiv:2104.01409*, 2021.
- [6] A. Triantafyllopoulos, B. W. Schuller, G. İymen, M. Sezgin, X. He, Z. Yang, P. Tzirakis, S. Liu, S. Mertes, E. André *et al.*, "An overview of affective speech synthesis and conversion in the deep learning era," *Proc. of the IEEE*, 2023.
- [7] Z. Du, B. Sisman, K. Zhou, and H. Li, "Disentanglement of Emotional Style and Speaker Identity for Expressive Voice Conversion," in *Inter-speech*, Sep 2022.
- [8] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.
- [9] H.-S. Oh, S.-H. Lee, D.-H. Cho, and S.-W. Lee, "Durflex-evc: Duration-flexible emotional voice conversion leveraging discrete representations without text alignment," *IEEE Tran. on Affective Computing*, 2025.
- [10] N. Raj Prabhu, N. Lehmann-Willenbrock, and T. Gerkmann, "In-the-wild speech emotion conversion using disentangled self-supervised representations and neural vocoder-based resynthesis," in *Proc. ITG Conf. on Speech Comm.*, Sep. 2023.
- [11] F. Kreuk, A. Polyak, J. Copet, E. Kharitonov, T. A. Nguyen, M. Rivière, W.-N. Hsu, A. Mohamed, E. Dupoux, and Y. Adi, "Textless speech emotion conversion using discrete & decomposed representations," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 2022.
- [12] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Int. Conf. on Learning Representations (ICLR)*, 2014.
- [13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [14] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," in *Advances in Neural Inf. Proc. Systems (NeurIPS)*, 2020.
- [15] N. Raj Prabhu, B. Lay, S. Welker, N. Lehmann-Willenbrock, and T. Gerkmann, "EMOCONV-Diff: Diffusion-based speech emotion conversion for non-parallel and in-the-wild data," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2024, pp. 11 651–11 655.
- [16] A. N. Salman, Z. Du, S. S. Chandra, İsmail Rasim Ülgen, C. Busso, and B. Sisman, "Towards naturalistic voice conversion: Naturalvoices dataset with an automatic processing pipeline," in *Interspeech 2024*, 2024, pp. 4358–4362.
- [17] F. Busquet, F. Efthymiou, and C. Hildebrand, "Voice analytics in the wild: Validity and predictive accuracy of common audio-recording devices," *Behavior Research Methods*, 2023.
- [18] K. Zhou, "Emotion modelling for speech generation," PhD thesis, National University of Singapore, 2022, available at <https://scholarbank.nus.edu.sg/handle/10635/243782>.
- [19] C. Robinson, N. Obin, and A. Roebel, "Sequence-to-sequence modelling of f0 for speech emotion conversion," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2019, pp. 6830–6834.
- [20] T.-H. Kim, S. Cho, S. Choi, S. Park, and S.-Y. Lee, "Emotional voice conversion using multitask learning with text-to-speech," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2020, pp. 7774–7778.
- [21] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Emotion intensity and its control for emotional voice conversion," *IEEE Tran. on Affective Computing*, 2023.
- [22] K. Zhou, B. Sisman, and H. Li, "Vaw-gan for disentanglement and recombination of emotional elements in speech," in *IEEE Spoken Language Tech. Workshop*, 2021.
- [23] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Tran. on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.
- [24] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 32, 2019.
- [25] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Ajiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Interspeech*, 2017, pp. 4006–4010.
- [26] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Ajiomyrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2018, pp. 4779–4783.
- [27] Z. Zeng, J. Wang, N. Cheng, T. Xia, and J. Xiao, "AlignTts: Efficient feed-forward text-to-speech system without explicit alignment," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*. IEEE, 2020, pp. 6714–6718.
- [28] W.-C. Huang, Y.-C. Wu, and T. Hayashi, "Any-to-one sequence-to-sequence voice conversion using self-supervised discrete speech representations," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2021, pp. 5944–5948.
- [29] S. Wang, T. Qi, C. Lu, Z. Luo, and W. Zheng, "Enhancing zero-shot emotional voice conversion via speaker adaptation and duration prediction," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2025.
- [30] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations," in *Interspeech*, 2021.
- [31] W.-N. Hsu, T. Remez, B. Shi, J. Donley, and Y. Adi, "Revise: Self-supervised speech resynthesis with visual input for universal and generalized speech regeneration," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2023.
- [32] D. de Oliveira, N. Raj Prabhu, and T. Gerkmann, "Leveraging semantic information for efficient self-supervised emotion recognition with audio-textual distilled models," in *Interspeech*, 2023.
- [33] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, M. Zeng, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2021.
- [34] L. I.-K. Lin, "A Concordance Correlation Coefficient to Evaluate Reproducibility," *Biometrics*, vol. 45, no. 1, p. 255, Mar. 1989.
- [35] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Tran. on Pattern Analysis and Machine Int.*, 2023.
- [36] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, 2008.
- [37] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, 1980.
- [38] N. Raj Prabhu, N. Lehmann-Willenbrock, and T. Gerkmann, "End-to-end label uncertainty modeling in speech emotion recognition using bayesian neural networks and label distribution learning," *IEEE Tran. on Affective Computing*, vol. 15, no. 2, pp. 579–592, 2024.
- [39] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: A unified framework for bandwidth extension and speech enhancement," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2023, pp. 1–5.
- [40] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, pp. 12 449–12 460, 2020.
- [41] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z.-H. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Inter-speech*, Apr 2018.